Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 3828 – 3835

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Capturing human motion in natural environments

Zhiqing Cheng[a], Anthony Ligouri[a], Ryan Fogle[b], Timothy Webb[c]

[a]*Infoscitex Corporation, 4027 Colonel Glenn Highway, Dayton, Ohio 45431, USA*
[b]*Wright State Research Institute, 4035 Colonel Glenn Highway, Dayton, OH 45431, USA*
[c]*Air Force Research Laboratory, 2800 Q Street, Wright Patterson Air Force Base, Dayton, OH 45433, USA*

Abstract

The problem of capturing human motion in a natural environment is discussed from the perspective of needs, significance, scenarios, and technical challenges. The technologies that can be potentially used to capture human motion and activity in a natural environment are briefly discussed with an emphasis on computer vision-based markerless motion capture technology. Three representative markerless motion capture methods for capturing human motion from video imagery are implemented in this paper. The synthetic data generated by modeling and simulation and the data collected in laboratory environments are used for the training and validation. The initial results of three methods are presented and analyzed, and the advantages and disadvantages of each are discussed and compared.

Keywords: Motion capture; Markerless motion capture; Computer vision; Discriminative model; Generative model; Pictorial structure

1. Introduction

Compared to traditional motion capture that is often conducted in a laboratory environment under controlled conditions, capturing human motion in natural environments with markerless motion capture (MMC) technologies can potentially provide great benefits which include:

- High biofidelity: Since human motion will be no longer restricted by the suits, markers, or other sensors being placed on the body, the captured motion may be more lifelike.
- True realism: Since humans move or act naturally, the captured motion could be more realistic.
- Motion variety: For instance, when human subjects wear loose clothing, it is almost impossible for marker-based methods to capture the true body motion.

doi:10.1016/j.promfg.2015.07.886

- Minimum pre-setting and no need for subject cooperation. These are two unique features provided by MMC technologies.

In this paper natural environments refer to any of the following conditions:

- Natural light, shadow, occlusions, natural terrain and backgrounds, and natural scenarios;
- Human in natural appearance wearing street clothing and possibly carrying some objects;
- Human in natural state performing actions/activities freely;
- Human in a natural context interacting with other humans or surroundings.

2. Needs and requirements

2.1. Needs

There are various industry needs and commercial scenarios where capturing human motion in natural environments becomes necessary, such as athletics and sports, health care, human machine interface, and entertainment. Human motion capture in natural environments has a variety of important applications. For instance, within the modeling and simulation (M&S) community, human activity M&S plays an important role in simulation-based training and virtual reality (VR). However, human motion/activity simulation provided by current human modeling tools/technologies is either artificially synthesized or based on data collected in a laboratory environment, which lacks sufficient biofidelity and realism. In order to describe and simulate human motion/activity in the real world, it is necessary to capture human motion in a natural environment. Another important application area is human centric intelligence, surveillance, and reconnaissance (ISR), where the captured motion can be fused with the data acquired from other sensors. The captured motion date may also serve as "ground truth" when human motion is to be analyzed or ascertained from other sensor data. For homeland security, human motion capture and analysis from video streams recorded in natural settings (e.g., airports and security check points) may help recognize human intention and identity human-borne threats.

2.2. Requirements

There are various performance criteria required for motion capture technology (MCT) to be used in natural environments based upon specific application scenarios. Major and common ones are as follows.

- *Accuracy* - With respect to different applications or scenarios, accuracy can be defined at three different levels.
    - Low level: The focus is on pose identification. The applications or scenarios include machine-human interface, human intention prediction, and human activity recognition.
    - Medium level: The emphasis is on pose identification as well as joint angle estimation. The problems can be, for example, human activity replication/animation in M&S based training and serious games where a trade-off between biofidelity and realism is required.
    - High level: The focus is on joint angle estimation and motion analysis. The applications involve biomechanics issues which require precise estimation of joint angles so that the relationship between force and motion can be accurately determined. The problems include sports (e.g., athletic training), health (e.g., prosthetic rehabilitation, and gait and balance training), and the extraction of spatial-temporal bio-signatures.
- *Efficiency* - Motion capture discussed in this paper includes recording and processing motion data. Recoding data is mainly related to sensors and relevant hardware, whereas processing data mainly utilizes software to derive the desired information, such as pose identification and joint angle estimation. Since extensive computational efforts may be needed in processing data, efficiency refers to the computational speed at which a designated task (e.g., pose identification) can be accomplished. Depending on specific tasks of motion data processing and applications, efficiency can be considered at two different levels.
    - Real-time: Ideally the processing of captured motion data can be done in real time or nearly real-time so that the desired data or information can be provided for timely use. For many application scenarios, such as

machine-human interface, immersive training, and security surveillance (e.g., human intention prediction and human-borne threat detection), it is necessary to achieve real-time processing.
  - Off-line: For many applications, such as those related to gait/motion analysis, activity replication, and bio-signature extraction, real-time processing is not necessary; instead, off-line processing is acceptable.
- *Robustness* - It has two fold implications: Sensors and hardware system can reliably acquire motion data under specified conditions; and the meaningful or desired data/information can be derived from the data collected. While it is desirable for the MCT to perform robustly for every frame, given the complexity of human motion under various natural conditions, it is almost inevitable that a system will fail for some ill-conditioned frames. These temporary failures may be acceptable, if the system can capture motion data for key frames reliably.
- *Minimum setting/interference* - Capturing human motion in natural environments often requires minimum system settings (e.g., setting lights, placing markers or sensors on a subject) or even prohibits pre-setting. In many application scenarios, such as security surveillance and human centric ISR, it is impossible to have subject's cooperation, and it is preferable to avoid subject awareness.

Due to the complexity of human motion and the variety of natural environments, capturing human motion in natural environments has great hurdles to overcome. From data collection perspective, the major technical challenge is acquiring reliable, useful, and complete data under various realistic/natural conditions. From data processing perspective, the main technical difficulty is quickly analyzing the data to derive the desired motion data or information.

3. State-of-the-art

The optical motion capture technology, as a gold standard in accuracy, is a widely used MCT and commercially provided by many vendors with various optical systems. The technology relies on line of sight between multiple cameras with light emitting strobes and retro-reflective markers placed on the subject. The optical systems are, however, cumbersome to move and cannot be used with common attire or street clothing, which inhibit its use in natural or real-world settings. Besides, using a marker tracking system is sometimes disadvantageous because of the mere fact that markers must be placed on the body. Not only do placing markers on the body introduce errors in the skeletal position due to soft tissue artifact, but markers often change the way subjects move, although the change is slight in most cases. For example, when markers are placed on the medial aspects of the arms and legs, some subjects will tend to walk bow-legged and with their arms out to avoid knocking off markers as the legs pass each other and the arms pass the torso.

The technologies that can be potentially used to capture human motion and activity in a natural environment include electromagnetic sensors, LED lights, inertial measurement units, range sensors (e.g., Microsoft Kinect), and vision-based MMC technology. Electromagnetic sensors provide accurate orientation and position, but are greatly limited by the range of the generated magnetic field. In recent years, depth cameras such as Kinect are available for full body motion tracking with reasonable prices. However, like optical motion capture systems, depth camera with inherent field of views, lead to very limited workspaces. Due to the design nature of these depth cameras and their use of Infrared (IR) sensors, the use of these systems in direct sunlight is problematic at best. Inertial measurement units (IMUs) could be used for human motion capture with great portability and flexibility. They would work almost everywhere, but are unable to maintain long-term accuracy because of sensor drifting and the interference from local magnetic fields. Vision based MMC approaches rely on image sequences from one or multiple cameras for human motion analysis. It could maintain long-term tracking with a certain degree of accuracy, but may often produce inaccurate results due to occlusion.

Markerless motion capture is motivated by the need in many situations to capture the motion of people or objects outside of a motion capture lab in a natural environment. It has great potential applications including interactive training, clinical gait analysis, surveillance, and vision for autonomous systems, among many others. Using computer vision methods to perform MMC from a sequence of video images has been a central topic for computer vision research in the last two decades and most methods can be grouped into three general categories:

discriminative model based approaches, generative model based approaches, and part based models [1, 2]. In this paper, one representative method from each category is investigated and implemented.

4. Technology development

4.1. Method-1: Discriminative method

The discriminative method implemented in this paper is based on the method proposed in [3] with two major modifications. One is using the discrete cosine transform (DCT) to replace shape context (SC) as the silhouette shape descriptor. While SC is an effective silhouette shape descriptor, it is calculated for every point of the silhouette contour for each frame, thus taking excessive computer time and memory when a large number of silhouettes are analyzed. The DCT is computed on the entire silhouette contour for each frame, thus taking much less computer time and memory. Then, all DCT coefficients or a truncation of its first part (the first 400 coefficients, for example) are used to form a DCT coefficient vector for the silhouette shape description. Another modification is that a principal component anlysis (PCA) is used to characterize the space formed by these DCT vectors. Then each DCT vector is projected onto the eigenspace formed by its principal components. Instead of directly using the DCT coefficient vector, the first 64 projection coefficients are used to describe the silhouette shape for each frame.

The silhouettes from nine subjects performing five activities simulated by [4] are used to train the method. Another subject (subject 1100) performing the same five activities is used as the test case. By comparing the estimated joint angles with their true values (which are from the BVH files used in the simulation) for subject 1100, the root mean square (RMS) error for each joint angle is calculated for walking and jogging and shown in Fig. 1. A particular pose of subject 1100, paired with the corresponding skeleton pose defined by the estimated joint angles, is shown in Fig. 2(a) and 2(b), respectively for walking and digging.

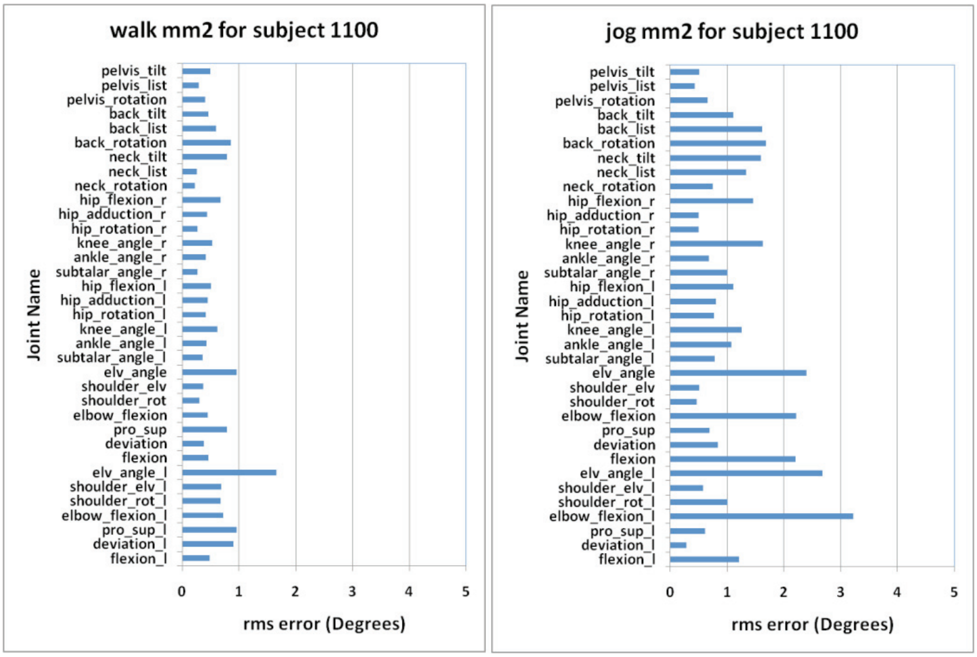

Fig. 1. RMS errors for walk and jog activities.

a b

Fig. 2. (a) Subject 1100 walking; (b) Subject 1100 digging.

4.2. Method-2: Generative method

The generative method is based on the method described in [5] with several modifications. One is that a full body model with a fixed shape rather than a torso model described by PCA coefficients is used. The cost function is also modified to reflect the fact that only the pose and position of the model (not shape) is considered in the optimization and a penalty factor evaluation is added to help overcome the problems with limb occlusion.

The animation videos of subject 1100 described above are used to create a contour distance image for each frame. It creates a pixel depth map with the lowest value, 0, along the outline of the silhouette, and larger positive values farther away from the outline, both inside and outside the silhouette figure. For each frame of video, there are two views: a head-on view aimed at the front of the subject, and an orthogonal view at 90 degrees to the subject's right. A three dimensional (3D) deformable human model is then created. In order to create the best possible fit, the model reflects the exact shape of subject 1100. The model has 15 bones and 36 degrees-of-freedom (DoF) that define the model's 3D pose, position, and orientation in space.

Using the covariance matrix adaptation evolution strategy (CMA-ES) algorithm by [6], the model parameters, i.e., joint angles and global 3D position, are optimized to best fit the silhouette. During optimization iteration, the model is positioned according to the guess for the present iteration. Front and side projections of the positioned model are then captured, creating similar silhouettes to those in the input video frame which are reduced to an image of the silhouette outline only using an edge finding algorithm. The length (number of pixels) in the model contour is calculated, and a penalty factor of 2 is applied if the contour lengths of the input image and model image are not similar. Otherwise, the penalty factor equals 1. The cost function is then calculated. Within the cost function, the front and side model outline images are overlaid on their respective contour distance images. For each view, the pixel values of the contour distance image are summed over the white pixels of the model contour and the sum is normalized by the number of pixels in the contour. The normalized sums of both views are added to calculate a total cost for the present iteration, which is multiplied by the penalty factor.

The method is tested with walking, jogging, and throwing motions performed by subject 1100. The time histories of optimal joint angles are smoothed using a second order Butterworth filter with a cutoff frequency of 6 Hz. Animations of the model using the filtered joint angles look smooth and quite realistic. An example of the input silhouettes (side and frontal), a contour distance image, and the model in the optimized position for a frame of walking motion is shown in Fig. 3.

The results of optimization are compared to the original BVH files used to create the test data described above to determine accuracy. Only major sagittal plane joint angles are compared, as they are the significant ones that are commonly analyzed in basic gait analysis. The RMS error for each joint angle for walking, jogging, and throwing is calculated. Of the three motions, walking has the smallest RMS errors. Over the length of the video, RMS errors for the lower limb joints and lumbar joint, i.e. lumbar flexion, hip flexion, knee flexion, and ankle plantar/dorsiflexion ranged from 4.2 to 14.2 degrees. The upper limbs experience much more occlusion from the torso; consequently, upper limb joint angle errors, i.e. shoulder and elbow flexion, are significantly larger, ranging

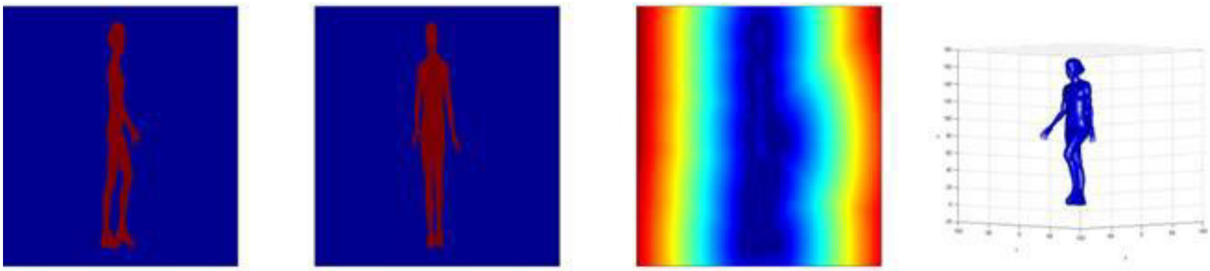

Fig. 3. Input silhouette, contour distance image, and optimized model in 3D pose.

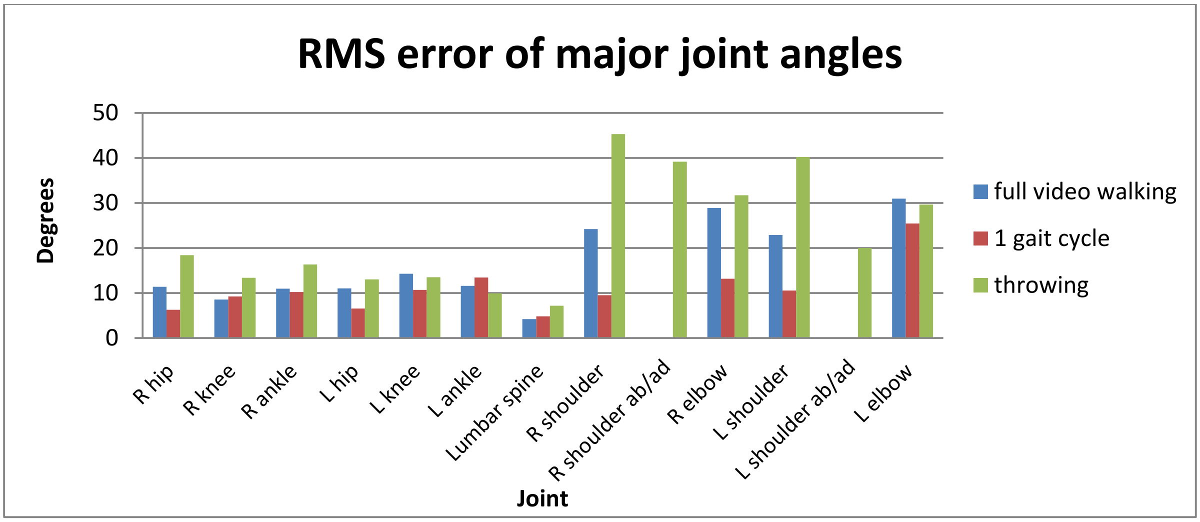

Fig. 4. RMS error of major joint angles during walking and throwing.

from 22.9 to 31.0 degrees. However, over the length of the video and multiple strides, limbs were occasionally confused by the model, which led to a solution with good fit but poor accuracy and increased RMS errors. Error calculations for one complete gait cycle without limb confusion showed smaller RMS errors. Fig. 4 shows a comparison of RMS errors for each joint angle between the full video and one complete gait cycle.

4.3. Method-3: Pictorial structure method

The pictorial structure method utilized in this study is based on human detectors described in [7]. The pictorial structure detectors are trained on truthed examples of objects – in this case, humans. These truthed images include boxes surrounding single body elements that may include the upper arms, lower arms, head, torso, lower leg, and upper leg. From the training data, histogram-of-oriented gradient (HOG) features are extracted and learned for each body parts. Additionally, the poses, or relative distances and directions between the various body parts, are learned as well. Intuitively, when applied to a test image, the detector searches the picture at multiple resolutions in an attempt to find regions with similar HOG features as the detector. The detections are refined and false alarms reduced by keeping only the regions that haves similar distances and orientations as the poses included in the training dataset. The result after application of the detectors is a two dimensional (2D) pose estimation in the form of skeleton-like connections, as seen in Fig. 5(a). The code for pre-trained detectors, training, and testing is freely available to the research community [8].

The pictorial structure method used above assumes a single 2D image. Oftentimes, however, multiple images of a human pose may exist. In such cases, the accuracy of the pose estimation may be further refined by applying part detectors to multiple images separately and then fusing them together in 3D space. The fusion occurs by taking a specific body part detection and back projecting (or tracing the ray path for a given image pixel in reverse direction of travel). Ideally, the back projected rays would intersect at the true joint location. Practically, though, the joint detections are inaccurate. These inaccuracies may result in noisy 3D estimates, or worse, rays that do not intersect. Because the rays are not guaranteed to intersect, the k-means algorithm is applied to obtain the appropriate number of centroids, or joint location estimates. As shown in Fig. 5(b), the result of this procedure is a 3D skeleton estimation.

The global coordinates of the joint centers found by the parts detector are then used to establish body-fixed, 3D coordinate systems for the pelvis, torso, upper arms, and thighs, respectively, with reference vectors pointing along the long axes of the head, lower arms, and shanks. For three DoF joints, such as the shoulder or hip joint, joint angles are found by calculating the Euler rotation angles between the 3D coordinate systems of the proximal and distal segments. For single DoF joints (e.g., knee and elbow), the angle between the distal reference vector and the coordinate that runs along the long axis of the proximal segment is given by the inverse cosine of the dot product of these two normalized vectors.

5. Discussion

Both the discriminative method and generative method implemented are able to track several human motions with varying degrees of accuracy. Both methods have advantages and disadvantages compared to the other. The discriminative method is more accurate in terms of RMS error for all joint angles and faster but requires training for the motions it expects to see. The method also has difficulty with motions that are not periodic. On the other hand, the generative method does not require any training but tracks motion with less accuracy. It can be seen from the results of both methods that the upper limb joint angles are tracked with greater error than the lower limbs. This is due to the occlusion by the torso. Whereas the lower limbs only occlude each other for a brief period of time, the upper limbs are often occluded by the torso for much longer periods of time, especially in the side view. When occluded, there is no information in the silhouette describing the position of the upper limbs. Therefore, many incorrect positions can have acceptable cost function values that may be accepted by the optimizer. The occlusion effects may be mitigated using multiple cameras to capture more views of the subject. Additional views would provide extra information that would reduce the ambiguity of the pose when limbs are occluded in other views.

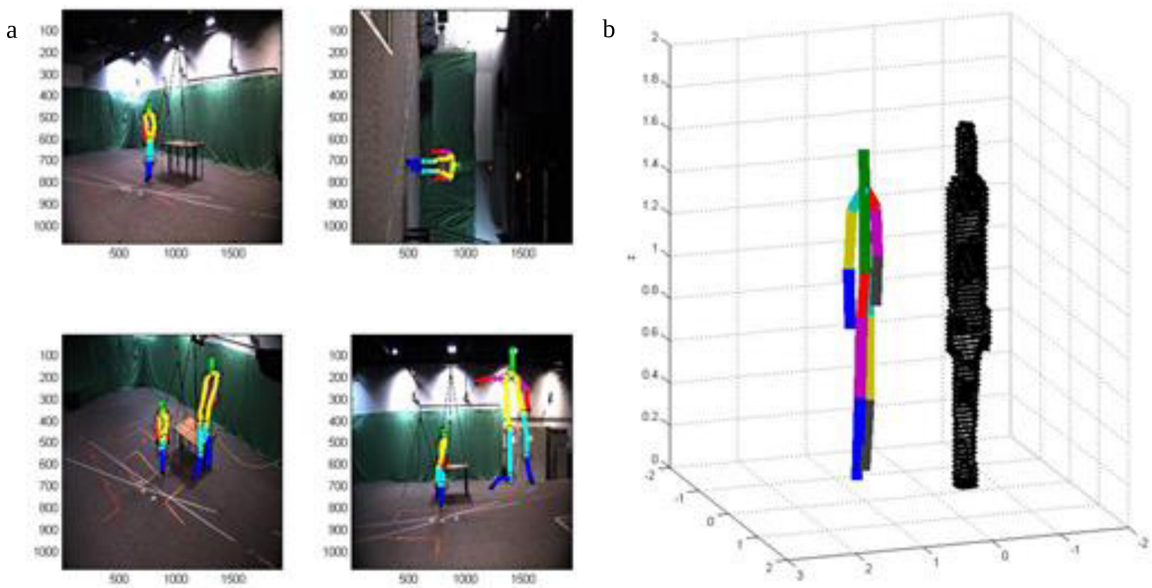

Fig. 5. (a) Part detectors applied to four camera perspectives; (b) 3D skeleton estimation from 2D pose estimation.

The performance of the pictorial structure method used in this paper has not been thoroughly tested and evaluated. Qualitatively, however, the algorithm has shown good performance for normal poses. For the poses with large joint angles or poses that are not included in the training dataset, the algorithm struggles. Additionally, in comparison to the 2D detections, the 3D detections appear more accurate and with fewer false detections. This is because the 3D joint locations are determined as a combination of the 2D estimates. Based on the limited tests, several key factors have been observed that could greatly increase the performance of the pictorial structure method:

- More unique camera view yield better results. In our tests six camera views vastly out-performs four views. Three camera views gave mediocre results.
- The detector needs trained on a larger variety of poses from many more view angles. The detector used in this paper did not include poses with the joints at any extreme angles, which made detecting motions other than standing, walking, or slow jogging difficult. Additionally, since cameras located directly overhead were not included in the training set, the potential advantage of such cameras in testing wasn't utilized. However, vertical cameras may eliminate much pose ambiguity.
- Image quality played a significant role in both the detector's ability to detect body segments and our background subtraction algorithm's ability to detect background. At times when the subject's skin or clothing blends with the background due to color similarity or shadow, both algorithms struggle.

6. Conclusions

Capturing human motion in natural environments provides various benefits and becomes necessary under various scenarios. However, due to the complexity of human motion and the variety of natural environments, it still has great hurdles to overcome before a reliable and effective technology becomes available for practical applications. Among the motion capture technologies that can be potentially used in natural environments, the computer vision based MMC technology has the largest potential, because it requires minimum pre-setting and has no need for subject cooperation. While different computer vision based methods have unique advantages, fusing or integrating these methods into a MMC technology may substantially increase its overall capability.

Acknowledgements

This study was carried out under the support of Small Business Innovation Research (SBIR) Phase II funding (FA8650-11-C-6226) provided by the U.S. Air Force.

References

[1] Leonid Sigal, Articulated Pose Estimation and Tracking, Chapter 8, Visual Analysis of Humans, ISBN: 978-0-85729-997-0, Springer 2011.
[2] Mykhaylo Andriluka, Leonid Sigal, and Michael Black, Benchmark Datasets for Pose Estimation and Tracking, Chapter 13, Visual Analysis of Humans, ISBN: 978-0-85729-997-0, Springer 2011.
[3] A. Agarwal and B. Triggs, Recovering 3D Human Pose from Monocular Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1):44-58, 2006.
[4] Z. Cheng, S. Mosher, J. Camp, and D. Lochtefeld, Human Activity Modeling and Simulation with High Bio-fidelity, Proceedings of 2012 I/ITSEC, Orlando, Florida, 2012.
[5] S. Saito, M. Kouchi, M. Mochimaru, and Y. Aoki, Model-based 3D human shape estimation from silhouettes for virtual fitting, Proceedings of SPIEI, 2014.
[6] N. Hansen, Cmaes.m, https://www.lri.fr/~hansen/cmaes.m, accessed March, 2014.
[7] Y. Yang and D. Ramanan, Articulated Human Detection with Flexible Mixtures of Parts, Proceedings of Computer Vision and Pattern Recognition, Colorado Springs, Colorado, 2011.
[8] D. Ramanan, Articulate Pose Estimation Code, Available at http://www.ics.uci.edu/~dramanan/software/pose/.